基于多特征融合的科技文献自动标引方法研究*

鞠佳辰；宋培彦

天津师范大学 天津 300387

摘要：[目的/意义]当前用户迫切需要在极度复杂的信息当中高效获取具有价值的信息，在这种背景下，本文提出一种多特征融合的自动标引方法以提高文本标引的准确性。[方法/过程]首先将文本正文和摘要同时作为标引源，接着分别采用 Keybert 方法和 TF-IDF 方法处理摘要和正文，同时结合统计学习法的词频特征和机器学习法的语义特征获取两组文本候选标引词；最后通过语义相似度计算做融合处理结合两种方法的优势以体现对标引结果的准确性和全面性的整体把握。[结果/结论]实验表明，基于多特征融合的文本自动标引是可行的，具有较好的标引结果。

关键词：自动标引 多特征融合 候选词提取

分类号：G353

1 引言

随着大数据时代的到来，用户迫切需要在极度复杂的信息当中高效获取具有价值的信息，解决信息资源无限增长和信息检索低下的矛盾。关键词是人们快速了解文档内容、把握主题的重要方式，被广泛应用于新闻、科技论文领域，以方便用户高效地管理和检索文档[1]。文本自动标引就是利用计算机系统模仿人的标引活动从拟存储、检索的事实情报或文献（题目、摘要、正文）中自动地标注出关键词的过程。与传统手工标引相比，自动标引具有处理能力强、处理效率高、成本低、一致性和稳定性高的优势，更能适应信息社会用户的检索需求[2]。随着网络时代信息的爆 炸式增长，文本自动标引已成为用户获取核心内容的重要手段。目前文本自动标引研究在标引准确性和全面性上仍在不断探索。

文本自动标引可分为自动抽词标引和自动赋词标引[3]。根据自动标引采用的理论依据来划分，自动标引可以分为统计分析方法、 语言分析方法、人工智能法和混合方法[3]。本文的研究方法是基于统计分析法和深度学习方法相结合的抽词标引方法。不管是统计学习法、语言分析法还是人工智能法存在各自优势的同时也有自身局限性。因此使用混合方法作为文

*本文系全国名词委 2020 年度科研项目“国际组织术语库集成方法研究”（项目编号：YB20200011）研究成果之一。

作者简介： 鞠佳辰，硕士研究生，E-mail：1157376259@qq.com 宋培彦，副教授，博士，E-mail：spyer2008@126.com，研究方向：知识组织、自然语言处理

本自动标引方法成为一种趋势。

综上所述，考虑到文本标引源的局限以及不同标引方法的优劣问题，本研究提出将文本正文和摘要同时作为标引源，并结合统计学习法的词频特征和机器学习法的语义特征获取文本候选标引词，再通过融合处理结合两种方法的优势以体现对标引结果的准确性和全面性的整体把握。

2 相关研究

有关文本自动标引的研究最早在国外盛行，卢恩[5]最先创立了以词频为特征的统计标引方法，又称词频统计标引法。我国对文献信息自动标引的研究开始于 20 世纪 80 年代初，起步比较晚，相较于国外的研究还有一定差距[6]。进入 21 世纪初，尤其是近年来相关研究逐渐丰富起来，自动标引技术基本达到实用水平。王小林[7]在 TF-IDF 的算法的基础上，将位置特征融入到算法，形成新的基于词频统计的关键词提取方法。姜艺、黄永[8]等人在传统的词频特征以及位置特征基础上，融合词汇功能特征，使用计算机领域的学术文献基于分类和排序两种思想进行关键词抽取实验。为了解决利用语义低效和抽取语义重复的弊端，有学者在 2018 年提出了一种具有注意力机制、复制机制和覆盖机制的序列到序列框架[9]。李千驹[10]等使用字符串模式匹配法，基于叙词表和关键词词表自动抽取关键词，在增量、组合以及排序方面有效优化了人工标引结果。国内在赋词标引的研究方面，王星，刘伟[11]提出了一种基于文献间引用关系，改进遗传算法，对学术文献进行自动标引的方法。章成志[12]从眼动特征的选择、眼动特征与文本特征的组合这两个方面，全面考察通用语料的眼动数据对微博关键词抽取任务性能的影响，同时提出了一个眼动数据的扩充方案用于解决眼动数据集与测试数据集在数据规模上相差较大这一问题。综上所述，我国对自动标引的研究正逐渐深入，尤其对标引方法的创新发展迅速。并且混合法作为自动标引的方法正逐渐受到重视，因此采取多特征融合思路这一混合标引方法作为文本自动标引方法是值得深入探索的。

3 基于多特征融合方法的自动标引模型

按照引入多特征的关键词提取基本思路，在标引模型提取关键词的过程中，尽可能融入更多特征关系来提高模型提取准确度。文本输入分为摘要类短文本和正文类长文本，将 KeyBERT 与 TF-IDF 算法优势互补，用 KeyBERT 算法处理摘要类短文本，TF-IDF 算法处理正文类长文本。并通过语义相似度计算实现标引词融合。因此，多特征融合方法的文本自动标引模型在确定文本双输入的情况下，经过融合处理单元计算，输出标引词表征文本核心内容。本文提出的多特征融合方法文本自动标引模型如下：

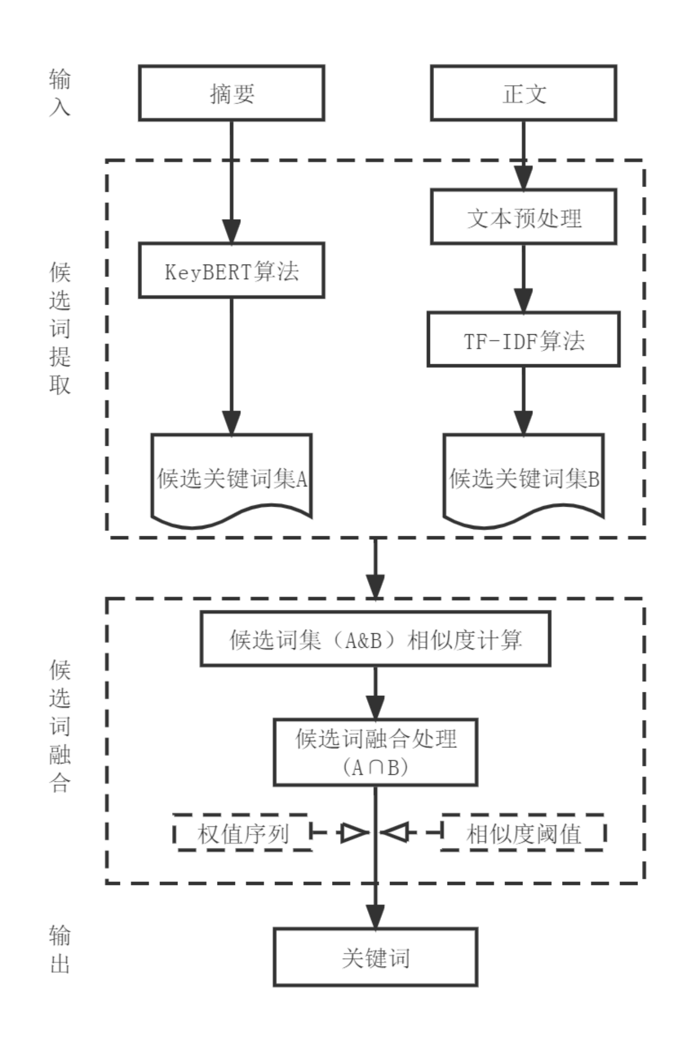


图 1 多特征融合方法主题自动标引模型

如上图所示，自动标引模型主要分为输入、候选词提取、候选词融合处理、输出四部分，依据信息重要程度与文本数量分为短文本和长文本，例如摘要是对正文内容的总结，关键词从摘要中提取出的概率将会更大，它的文本数量虽然少但是信息重要程度较高，因此归类于短文本；而正文内容归类于长文本。候选词提取部分负责输入文本候选词的初步筛选，分别对应生成候选词集合，KeyBERT 算法前期包括了分词、去停用词等预处理，所以在模型中并未单独列出。候选词融合处理部分负责将输入文本的候选词进行融合筛选，候选词融合处理单元核心采用余弦相似度原理，其结构组成如下：

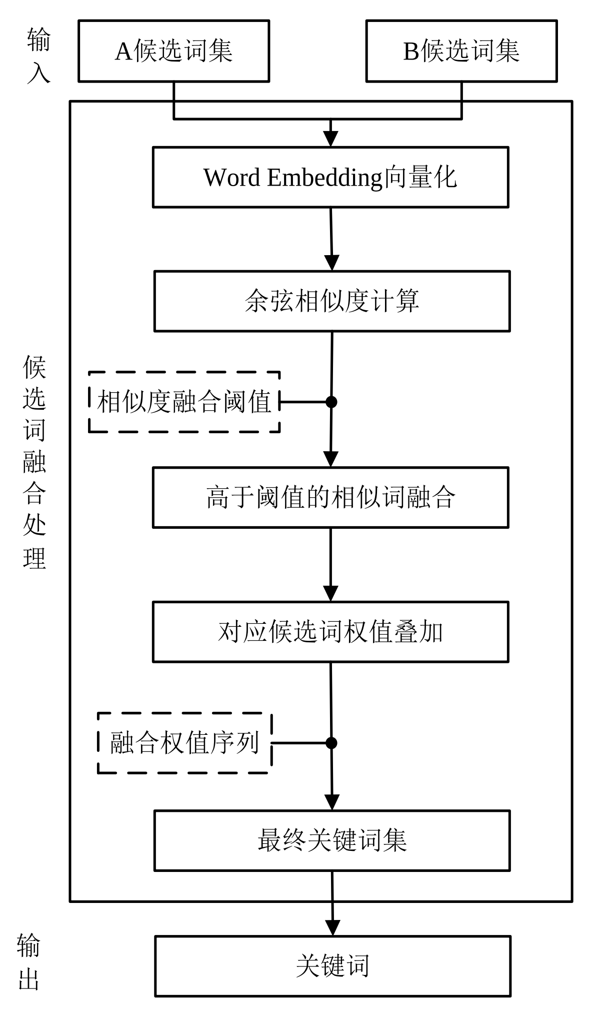


图 2 候选词融合处理单元

对不同算法和不同文本生成候选词集进行融合处理，候选词集本身与候选词集之间部分词汇会具有较强的凝聚力，其含义具有相似性，往往相似性越高的词汇越能承当起主题关键词，将相似性高的词汇提取并融合是候选词融合处理模块的核心工作。如上图所示，实线框是融合处理单元结构，虚线框是控制条件，经过词集间遍历互融合，在相似度融合阈值确定下，高于阈值的词汇将被筛选出作为候选词，再经过最终融合权值序列条件，排序前列的候选词将被确定为最终关键词集。改变相似度融合阈值和融合权值序列会对模型计算处理结果产生微调，后续可通过实验方式取得较好的设置参数。

多特征融合方法主题自动标引模型能够从文本不同角度出发，既考虑了摘要类短文本的重要程度，也没有忽略正文长文本中体现出的关键信息。它结合了 TF-IDF 高效性，并将语义特征融合，弥补了上下文联系不足的缺陷。

4 关键词抽取算法

4.1TF-IDF 关键词抽取

TFIDF（Term Frequency&Inverse Documentation Frequency）算法是 Stilton[5]提出的，主

要思想是：如果一个词在特定的文档中出现的频率越高，TF 值越大，代表它表达该文档内容的能力越强，应该被赋予较高的权重；如果一个词在一组文档中出现的范围越小，计算得到的 IDF 值越小，说明它区分文档内容的能力越强，应该被赋予较高的权重。

TF-IDF 的计算公式如下所示：

$$TF-IDF(t,d)=\frac{TF(t,d)}{DF(t)}=TF(t,d)\bullet IDF(T)$$

以上公式中，t 代表单词（term），d 代表文档（document），TF（t，d）代表单词 t 在文档 d 中的出现频次，DF（t）代表包含单词 t 的文档数，DF 的倒数就是 IDF。由此可见，TF-IDF 模型主要利用统计学原理获取在当前文本中出现频率高而在其他文档中出现频率低的词语，即能够代表文档特色的词语。

TFIDF 算法的优点很明显：首先，TFIDF 算法的原理简单且容易实现；其次，它相对全面地考虑到特征项在单个文本中和在文本集中的情况，经 TF 和 IDF 双重选择得到的特征项更具代表性。

4.2KeyBERT 关键词抽取

KeyBERT 是一种新型且简单的关键字抽取技术，其原理是利用 BERT 嵌入来创建与文档最相似的关键词和关键字短语。BERT 模型最终的目的是要使用无标注语料训练来获得文本之中的语义信息，简单的来说就是文本所具 有的语义表示，之后将语义表示在某个特定的 NLP（Natural Language Processing）任务中作微调，最终应用于该 NLP 任务。在基于深度神经网络的 NLP 方法中，文本中的词通常都用一维向量来表示（一般称之为“词向量”）。因此，对于BERT模型之中的核心信息输入主要是指原始词的一些向量，对于此向量可以进行初始化，而且也可以利用类似Word2vec的一些算法来进行训练；其中的信息输出主要的含义是文本里的词所包含的语义信息，所使用的向量表示。

关键词与文档在语义表示上是一致的，利用 BERT 的编码能力，能够取得较好的结果。但是缺点也很明显，首先，不同的语义编码模型会产生不同的结果；另外，BERT 只能接受限定长度的文本，使得在处理长文本时需要进一步先提取摘要等预处理措施，增加了时间复杂度，降低了准确率，因此，本文将 KeyBERT 算法应用于摘要文本的关键词提取，具有一定的针对性和实用性。

4.3 语义相似度计算

本文从不同的文本角度出发，对摘要和全文文本有针对性的进行关键词提取处理，采用不同的适应性算法提取出不同的关键词集合，融合多种语言特征作为提取准则，较为全面考

虑到不同特征对关键词提取的影响。本文提出的融合处理算法核心是语义相似度计算，对不同关键词集进行相似度处理形成新的关键词集，新的关键词集能够担任核心词的角色，成为表征文章内容的标签。关键词集的相似度处理过程与文本相似度计算过程类似，分词过程已在文本预处理阶段完成，且提取出的候选关键词集已具有代表性，直接对其进行向量化处理即可。由于目前缺乏高覆盖度的词汇知识库，因此采用语义向量捕获词汇关系。

Word2vec 在训练词向量中可以根据给定的语料信息将每个词汇向量化，在此过程中优化内部的训练模型机制，实现 word embedding，是目前主流的词汇向量化手段，它的主要模型包括 CBOW 模型和 Skip-gram 模型。对向量化后的词汇进行相似度计算，本文采用余弦相似度计算词汇向量间的相关性，余弦相似度用向量空间中两向量夹角的余弦值作为衡量两个个体之间差异的大小。

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

余弦值越接近 1，表明两个向量的夹角越接近 0 度，则两个向量越相似，反之，当余弦值越接近于 0，表明两个向量的夹角越接近 90 度，两个向量越不相似。在案例文本通过构建空间向量表示为两个向量 A、B 后，可以通过计算向量相似度来衡量两个词汇的语义相似度。

5 实验与分析

5.1 文本预处理

获取数据：为了保证模型的顺利构建以及模型输出的准确性，实验数据需选取一篇科技文献且该文献必须包含摘要、关键词和正文部分。本文实验以科技文献中的自动化领域的一篇文章为例进行实验，选取此篇文章的主要原因是：一是文章专业术语较多，便于开展学习训练和语义结合；二是文本主题特点显著，便于捕捉词向量特征。

文本预处理：将采集的实验文本进行切分，对实验文本正文部分进行预处理操作，包括分词、词性标注和停用词过滤；文本摘要部分可不单独进行预处理（KeyBert 模型封装了分词等预处理操作）。

以一篇自动化领域科技文献为例（以下实验均以此文献为例），对其主要研究内容进行分词和词性标注处理，其结果如下。

原句子为：针对传统程序设计方法难以对具有并发、协调、竞争等复杂事件的PLC控制系统编程的问题，探讨了一种基于Petri网模型的PLC程序设计方法，该方法利用Petri网的图形性质和对并发事件建模的能力，可简便、直观地构建PLC控制系统程序，并能对控制系统性能进行分析，以确认程序的合理性。

图 3 原句输出结果

分词后：针对/传统/程序设计/方法/难以/对/具有/并发/、/协调/、/竞争/等/复杂/事件/的/PLC/控制系统/编程/的/问题/，/探讨/了/一种/基于/Petri/网/模型/的/PLC/程序设计/方法/，/该/方法/利用/Petri/网/的/图形/性质/和/对/并发/事件/建模/的/能力/，/可/简便/、/直观/地/构建/PLC/控制系统/程序/，/并/能/对/控制系统/性能/进行/分析/，/以/确认/程序/的/合理性/。

图 4 分词输出结果

对上述分词后的科技文献段落进行去停用词处理，得到的结果如下：

去除停用词后：传统/程序设计/方法/难以/具有/并发//协调/竞争/复杂/事件/PLC/控制系统/编程/问题/探讨/一种/基于/Petri/网/模型/PLC/程序设计/方法/方法/利用/Petri/网/图形/性质/并发/事件/建模/能力/简便/直观/构建/PLC/控制系统/程序/能/控制系统/性能/进行/分析/确认/程序/合理性

图 5 去停用词输出结果

结果显示，利用 jieba 第三分词库可以过滤掉绝大部分的停用词与无效词，这些介词与连接 词不但会增加自动标引的工作量，也会干扰到标引精度，经过有效的数据清洗工作，得到较为清晰的数据材料以供后续文本标引使用。

5.2 关键词提取

文本摘要部分已经是文本的重点内容，使用 keyBert 算法直接设置抽取权重前 N 个词语作为抽取出候选关键词集合 A。

文本正文部分经过文本预处理后生成一组候选词集合，然后使用 TF-IDF 算法逐词遍历候选词集合，每个词语节点的最终得分进行由大到小排序，抽取得分高的前 N 个词语，作为候选关键词集合 B。

按照实验思路，对此篇论文的摘要和正文分别进行标引词提取实验，其中，候选标引词个数阈值统一设置为 10，得到的结果如下：

表 1 实验 1-摘要 Keybert 处理结果

序列	候选词	权重	序列	候选词	权重
1	程序设计	0.6431	6	编程	0.3919
2	控制系统	0.5610	7	模型	0.3891
3	Petri	0.5368	8	程序	0.3872
4	PLC	0.4811	9	性质	0.3804
5	图形	0.3965	10	性能	0.3787

表 2 实验 2-正文 TF-IDF 处理结果

序列	候选词	权重	序列	候选词	权重
1	Petri	0.6459	6	状态	0.5046
2	PLC	0.6215	7	事件	0.4846
3	token	0.5846	8	系统	0.4372
4	程序	0.5746	9	控制	0.3432
5	模型	0.5546	10	库所	0.2429

对比上述对文章结构不同的处理方法结果，可大致看出结果存在共性，这是因为摘要已经是对全文的总结，摘要中出现的词语大概率将会是全文的主题，而这些词将会在正文中重复出现体现研究主旨，前者语义特征捕捉较为精确的总结性质的短文本，后者通过 TF-IDF 捕捉词频特征的全部正文。不同候选词集具有共性和相似性才有必要进行下一步融合处理。

5.3 关键词融合处理

对于已经生成的两组关键词集合 A 和 B 做融合处理并生成最终的关键词集合 C 作为最终的关键词集。

（1）候选词的特征权值计算。

参数设置：因为这里是个层层推进数值对比试验，所以参数的设置不需要通过严格试验取得，给它们赋一定值并不影响试验结果的对比，只要保持同一参数始终保持同一值即可。Keybert 对摘要的标引词结果记为 A 集，对应权值记为 Qa；改进的 TF-IDF 对文本标引词结果记为 B 集，对应权值记为 Qb。融合处理与权值计算见下式。

$$A = \{a_1, a_2, a_3, ..., a_x\}$$

$$Qa = \{Qa_1, Qa_2, Qa_3, ..., Qa_x\}$$

$$B = \{b_1, b_2, b_3, ..., b_y\}$$

$$Qb = \{Qb_1, Qb_2, Qb_3, ..., Qb_y\}$$

上式中，a_x 表示不同的 Keybert 候选标引词，Qa_x 为对应的权值，x 为 A 集候选标引词个数；b_x 表示不同的改进 TF-IDF 候选标引词，Qb_y 为对应的权值，y 为 B 集候选标引词个数。

Qa 与 Qb 已进行归一化处理，取值区间均为(0,1)，因此可进行权值叠加处理。相似度融合界限值为 α，两词间相似度高于或等于 α 值定义为相似，且可进行融合处理。假设 A 标引词集 x 为 4，B 标引词集 y 为 6，标引词 a_1 与 b_2 两者相似，标引词 a_2 与 b_5、b_6 三者相

似，则融合处理后的标引词集 C 如下。

$$C = \{c_1 = a_1, c_2 = a_2, c_3 = a_3, c_4 = a_4, c_5 = b_1, c_6 = b_3, c_7 = b_4\}$$

$$\begin{aligned} & Qc = \{Qc_1 = Qa_1 + Qb_2, \\ & Qc_2 = Qa_2 + Qb_5 + Qb_6, \\ & Qc_3 = Qa_3, Qc_4 = Qa_4, \\ & Qc_5 = Qb_1, Qc_6 = Qb_3, Qc_7 = Qb_4\} \end{aligned}$$

（2）合并与排序。

因 A 集为摘要类型标引词，其词特征本身具有较高规范性，其表达含义更能体现科技文献主旨信息，所以此处候选词融合过程中 A 集优先级高于 B 集优先级，作并处理。将部分包含重叠的候选词合并将它们的权重值相加，最后将所得的每个候选词的值按降序排列。

（3）输出标引词。

从降序排列的候选词中抽取前 n 个词，作为最终的关键词输出。最关键的参数是相似度界限α，即定义多大的α确定为候选词相似，综合测试$\alpha = 0.8$时融合效果较好，限制标引结果序列为 5，即融合处理后排名前 5 的词作为最终标引词集，结果如下。

$$\text{标引结果}C\text{集} = \{\text{程序设计、控制系统、}Petri\text{、}PLC\text{、模型}\}$$

对照作者给出的参考关键词集如下。

$$\text{参考关键词集} = \{\text{Petri网、}PLC\text{、控制模型、程序设计、逻辑方程}\}$$

5.4 结果分析

以文献作者列出的关键词作为参考关键词，分析结果可知标引模型的输出结果与参考关键词集有较高的相似性，标引效果较好。实验结果表明，通过本文所提出的多特征融合的自动标引模型对抽取出的两组关键词组融合处理可以得到较为准确的输出结果，因此基于多特征融合的文本自动标引方法是可行的，标引准确率较高。

6 结语

本文根据多特征融合的基本思路，提出了基于多特征融合方法的自动标引模型。该模型分为输入、候选词提取、候选词融合处理和输出四部分，其中候选词提取分别采用 Keybert 方法混和 TF-IDF 方法处理摘要和正文并提取出两组候选关键词；在候选词融合处理部分核心技术采用余弦相似度计算。该模型在一定程度上既集合了两种算法的优点，又综合考虑到文本标引的准确性和全面性，对于文献自动标引关键信息提供了一种可行思路，具有一定的应用价值。

参考文献：

[1]韩红旗，桂婕，张运良，翁梦娟，薛陕，悦林东．大规模主题词自动标引方法[J]．情报学报, 2022, 41(05): 475-485.

[2]余春．自动标引研究进展[J]．图书馆学研究, 2012, (04): 18-22.

[3]蔡迎春，赵心如，朱玉梅，汪秀秀．我国文献标引技术的回顾与展望[J]．图书馆杂志, 2022, 41(03): 18-31.

[4]张静．自动标引技术的回顾与展望[J]．现代情报, 2009, 29(04): 221-225.

[5]Hans Peter Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal of Research and Development, 1957, 1(4).

[6]赵捧未，窦永香等．信息资源管理技术[M]．北京：科学出版社. 2010.

[7]王小林．改进的 TF-IDF 关键词提取方法[J]．计算机科学与应用, 2013, (1): 64-68.

[8]姜艺，黄永，夏义堃，李鹏程，陆伟．学术文本词汇功能识别——在关键词自动抽取中的应用[J]．情报学报, 2021, 40(02): 152-162.

[9]Zhang Y, Xiao W. Keyphrase Generation Based on Deep Seq2seq Model[J]. IEEE ACCESS, 2018, 6: 46047－46057.

[10]李千驹，李思达，刘建毅．一种基于知识组织的关键词自动标引方法[J]．情报科学, 2016, 34(11): 107-110+139.

[11]王星，刘伟．基于引文的中文学术文献自动标引方法研究[J]．图书情报工作，2014, 58(03): 106-110+105.

[12]章成志，胡少虎，张颖怡．通用语料的眼动数据对微博关键词抽取的性能提升探究[J]．情报学报, 2021, 40(04): 375-386.

[13]G. Salton, A. Wong, C. S. Yang. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11).

作者贡献说明：鞠佳辰：提出研究思路，论文写作；宋培彦：研究框架设计，论文修改

Automatic indexing of scientific and technological documents based on multi-feature fusion

JuJiachen;SongPeiyan

Tianjin Normal University　Tianjin　300387

Abstract： [Purpose/Significance]With the advent of the Big Data era, users are in urgent need of efficient access to valuable information in the midst of extremely complex information, especially in literature reading, where it is crucial to quickly grasp the core content and topic ideas of the text. [Method/Process] This study proposes to use both text body and abstract as citation sources, and combine the word frequency features of statistical learning method and semantic features of machine learning method to obtain text candidate citation words, and then combine the advantages of both methods by semantic similarity calculation to reflect the accuracy and comprehensiveness of the citation results as a whole.[Result/Conclusion] The experiments show that automatic text citation based on multi-feature fusion is feasible and has better citation results.

Keywords： automatic indexing   multi-feature fusion   candidate word extraction